บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ผู้วิจัยเสนอแนวคิดและทฤษฎีที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันโดยแบ่งการนำเสนอออกเป็น 5 ตอนคือ

ตอนที่ 1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีทฤษฎีตอบสนองข้อสอบ

ตอนที่ 3 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันด้วยวิธีถดถอยโลจิสติก

ตอนที่ 5 งานวิจัยที่เกี่ยวข้องกับการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

ตอนที่ 1 ความเป็นมาของการทำหน้าที่ต่างกันของข้อสอบ

การศึกษาเรื่องข้อสอบทำหน้าที่ต่างกันสำหรับกลุ่มผู้สอบย่อยมีมานานแล้ว เดิมเรียกว่าความลำเอียงของข้อสอบ (item bias) หรือการทดสอบความยุติธรรมของข้อสอบ ซึ่งมีการศึกษาอย่างจริงจังในช่วงปลายทศวรรษ 1960 มีการนำเสนอวิธีการต่างๆเพื่อนำไปใช้ศึกษาความแตกต่างทางวัฒนธรรมในการทำข้อสอบของนักเรียนผิวดำและนักเรียนผิวขาว สำหรับระบุข้อสอบที่ไม่เหมาะสมต่อผู้สอบกลุ่มย่อยและตัดข้อสอบออกจากแบบสอบ ขณะเดียวกันได้มีการศึกษาแบบสอบที่ใช้ในการคัดเลือกผู้สอบไม่ว่าจะเป็นการคัดเลือกเพื่อการศึกษาต่อ เพื่อบรรจุเข้าทำงาน เพื่อเลื่อนตำแหน่งหรืออื่นๆ ผลพบว่าแบบสอบไม่เป็นไปตามสัดส่วน ระดับสติปัญญาและเกิดความลำเอียงต่อผู้สอบกลุ่มย่อย จึงทำให้เกิดความจำเป็นต้องพัฒนาวิธีการตรวจสอบความลำเอียงของข้อสอบเพื่อนำข้อสอบไปปรับปรุงหรือตัดข้อสอบที่ลำเอียงต่อผู้สอบกลุ่มย่อยออกจากแบบสอบ ดังนั้นการตรวจสอบความลำเอียงของข้อสอบจึงเป็นส่วนหนึ่งของการพัฒนาแบบสอบและประเมินแบบสอบ (Angoff, 1993) ตัวอย่างเช่น ศูนย์บริการทดสอบทางการศึกษา (Educational Testing Service) ของประเทศสหรัฐอเมริกา ได้มีการตรวจสอบความลำเอียงทั้งก่อนและหลังจากการนำแบบสอบไปทดลองใช้ การตรวจสอบก่อนนำไปทดลองใช้จะใช้การตรวจสอบโดยผู้เชี่ยวชาญซึ่งประกอบด้วยบุคคลหลายฝ่ายตรวจสอบความไว (sensitivity) ของแบบสอบโดยพิจารณาถึงรูปแบบของข้อสอบ เนื้อหา ภาพประกอบ คำที่ใช้และอื่นๆ เพื่อไม่ให้เกิดความลำเอียงต่อผู้สอบกลุ่มย่อยกลุ่มใดกลุ่มหนึ่ง (กาญจนา วัฒนสุนทร, 2537)

ในระยะหลังได้มีการศึกษาเรื่องความลำเอียงของข้อสอบกันอย่างกว้างขวาง จึงทำให้เกิดความสับสนในการใช้คำและความหมาย เช่น ความลำเอียงของข้อสอบเป็นการตัดสินข้อสอบเพื่อนำไปสู่จุดประสงค์ของแบบสอบ หรือเป็นความสัมพันธ์กับประสบการณ์ของผู้สอบกลุ่มย่อยในการทำข้อสอบ และเป็นสารสนเทศทางสถิติที่ได้จากข้อสอบ เป็นต้น ดังนั้นเพื่อคลายความสับสน นักวิจัยส่วนใหญ่จึงใช้สารสนเทศทางสถิติที่ตรวจสอบได้มาเป็นเกณฑ์ในการตัดสิน และวิธีการใหม่ๆที่ใช้ตรวจสอบความลำเอียงนั้นจะเน้นไปที่ความแตกต่างระหว่างกลุ่มผู้สอบที่ตอบสนองต่างกันต่อข้อสอบข้อเดียวกัน ความแตกต่างกันนี้อาจมาจากข้อคำถาม ประสบการณ์หรือพื้นฐานเดิมที่แตกต่างกัน ในบางสถานการณ์ก็ไม่เหมาะสมที่จะใช้คำว่าลำเอียง ดังนั้นจึงควรเปลี่ยนมาใช้คำว่าข้อสอบทำหน้าที่ต่างกันซึ่งเป็นคำที่มีความเป็นกลางและเหมาะสมกว่า (Holland and Thayer, 1988 ; Angoff, 1993) นอกจากนี้ Camilli และ Shepard (1994) เสนอแนวคิดว่าการทำหน้าที่ต่างกันของข้อสอบเน้นที่วิธีการทางสถิติที่นำมาตรวจสอบข้อสอบทำหน้าที่ต่างกันจากผู้สอบต่างกลุ่ม ส่วนความลำเอียงของข้อสอบถ้าเน้นที่วิธีการตรวจสอบทางสถิติเพียงอย่างเดียวยังสรุปไม่ได้ว่าข้อสอบลำเอียง ต้องรวมถึงการวิเคราะห์เชิงตรรก (logical analysis) โดยอาศัยผู้เชี่ยวชาญพิจารณาเนื้อหาสาระของข้อสอบและจุดมุ่งหมายในการวัดแบบสอบก่อนที่จะระบุว่าข้อสอบนั้นลำเอียงหรือไม่ ดังนั้นข้อสอบที่ระบุว่าลำเอียงจึงเป็นกลุ่มข้อสอบย่อย (subset) ของข้อสอบที่ทำหน้าที่ต่างกัน และปัจจุบันพบว่านักวิจัยส่วนใหญ่ใช้คำว่าการทำหน้าที่ต่างกันของข้อสอบ ซึ่งมีผู้ให้ความหมายไว้ดังนี้

Kederman และ Macready (1990) กล่าวว่าการทำหน้าที่ต่างกันของข้อสอบ หมายถึงคะแนนข้อสอบที่ได้จากกลุ่มผู้สอบที่มีความสามารถเท่ากันแต่มาจากต่างกลุ่มกันมีความแตกต่างกันอย่างมีระบบ

Camilli และ Shepard (1994) กล่าวว่าการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ เป็นการตรวจสอบความเป็นพหุมิติของข้อสอบ โดยจะแสดงการทำหน้าที่ต่างกันระหว่างกลุ่มผู้สอบตั้งแต่สองกลุ่มขึ้นไปที่มีความสามารถหลัก (primary abilities) เท่ากัน แต่มีความสามารถรอง (secondary abilities) แตกต่างกัน

Potenza และ Dorans (1995) กล่าวว่าการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผลการตอบข้อสอบระหว่างกลุ่มผู้สอบสองกลุ่มที่นำมาเปรียบเทียบมีความแตกต่างกัน การเปรียบเทียบกลุ่มผู้สอบเป็นสิ่งสำคัญที่จะอธิบายถึงความแตกต่างระหว่างการทำหน้าที่ของข้อสอบกับคุณลักษณะแฝงของกลุ่มผู้สอบ

Narayanan และ Swaminatan (1996) กล่าวว่าการทำหน้าที่ต่างกันของข้อสอบ หมายถึง ผู้สอบมีความสามารถระดับเดียวกันแต่มาจากกลุ่มย่อยต่างกัน มีโอกาสในการตอบข้อสอบได้ถูกต้องแตกต่างกัน

ดังนั้นจึงสรุปได้ว่า การทำหน้าที่ต่างกันของข้อสอบ เกิดขึ้นเมื่อนำข้อสอบหรือแบบสอบไปทดสอบกับผู้สอบที่มีความสามารถหลักระดับเดียวกันหรือมีคุณลักษณะแฝงที่ต้องการวัดเท่ากัน แต่มีคุณลักษณะแฝงอื่นแตกต่างกัน ทำให้ผู้สอบต่างกลุ่มที่นำมาจับคู่เปรียบเทียบมีโอกาสตอบข้อสอบถูกแตกต่างกัน

ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบดำเนินการโดยเปรียบเทียบผลการตอบระหว่างผู้สอบ 2 กลุ่ม คือกลุ่มอ้างอิง (reference group) และกลุ่มเปรียบเทียบ (focal group) กลุ่มอ้างอิงเป็นกลุ่มที่คาดว่าจะได้ประโยชน์ในการตอบข้อสอบคือมีโอกาสในการตอบข้อสอบถูกได้มากกว่าผู้สอบอีกกลุ่มหนึ่ง ส่วนกลุ่มเปรียบเทียบเป็นกลุ่มที่สนใจศึกษาเป็นกลุ่มที่คาดว่าจะเสียประโยชน์ในการตอบข้อสอบ เช่น การศึกษาการทำหน้าที่ต่างกันของข้อสอบระหว่างผู้สอบต่างเชื้อชาติ กลุ่มเปรียบเทียบได้แก่ กลุ่มผู้สอบผิวดำ ในขณะที่กลุ่มอ้างอิงได้แก่ กลุ่มผู้สอบผิวขาวเป็นต้น ในการเปรียบเทียบจะศึกษาปัจจัยอันเกิดจากผู้สอบซึ่งส่งผลให้เกิดการได้ประโยชน์และเสียประโยชน์ระหว่างกลุ่มผู้สอบเช่น เพศ สีผิว เชื้อชาติ ภาษา สถาบันการศึกษา ประสบการณ์ เป็นต้น ต่อมาระยะหลังได้มีการศึกษาเปรียบเทียบวิธีการต่างๆในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบทั้งนี้เพราะมีวิธีการตรวจสอบหลายวิธีที่ถูกคิดค้นและพัฒนาปรับปรุง เพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันได้อย่างมีประสิทธิภาพมากที่สุด

วิธีการในการตรวจสอบการทำหน้าที่ต่างกันมีหลายวิธี ทั้งนี้เพราะมีการศึกษาและคิดค้นวิธีการต่างๆเพื่อให้สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบได้อย่างมีประสิทธิภาพมากที่สุด ซึ่งสามารถแบ่งตามประเภทการวิเคราะห์ได้ดังนี้ (Potenza and Dorans, 1995)

วิธีแมนเทล-แฮนส์เซล (Mantel-Haenszel : MH) เป็นวิธีที่พัฒนามาจากไคสแควร์แบบดั้งเดิม ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถโดยวิเคราะห์ทีละระดับความสามารถ มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและการทดสอบนัยสำคัญทางสถิติ เป็นวิธีที่มีความสอดคล้องกับวิธี IRT อีกทั้งสามารถใช้วิธี MH แทนวิธี IRT ได้อย่างประหยัดกว่า (Hambleton et al., 1986 อ้างถึงใน กาญจนา วัฒนสุนทร, 2537) ข้อดีของวิธีนี้ คือคำนวณง่าย ใช้ได้กับกลุ่มตัวอย่างขนาดเล็ก

วิธีถดถอยโลจิสติก (logistic regression : LR) เป็นวิธีการที่มีโมเดลพื้นฐานทางสถิติจึงสามารถตรวจสอบการทำหน้าที่ต่างกันของแบบสอบได้ถูกต้องใกล้เคียงกับวิธี IRT และเข้าใจธรรมชาติการทำหน้าที่ต่างกันภายในแบบสอบได้ดี วิธีนี้มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของแบบสอบแบบเอกรูป (uniform DIF) ได้เท่าเทียมกับวิธีแมนเทล-แฮนส์เซล และวิธีซิบเทสท์ (Swaminatan and Roger, 1990 ; Ackerman, 1992) นอกจากนี้ยังสามารถตรวจสอบการทำหน้าที่ต่างกันของแบบสอบแบบอเนกรูป (nonuniform DIF) ได้ดีเช่นเดียวกัน

วิธีทำให้เป็นมาตรฐาน (standardization : STND) โดยพื้นฐานแล้ววิธีนี้เป็นวิธีเชิงบรรยายไม่มีการทดสอบนัยสำคัญ หลักการสำคัญคือเป็นการเปรียบเทียบการถดถอยระหว่างคะแนนข้อสอบกับคะแนนแบบสอบของผลการตอบข้อสอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ มีหลักการวิเคราะห์การทำหน้าที่ต่างกันคล้ายกับวิธีแมนเทล-แฮนส์เซล ตรงที่ใช้หลักความแตกต่างระหว่างสัดส่วนการตอบข้อสอบที่ถูกที่ควรจะเป็นกับที่สังเกตได้ระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบให้ดัชนีการทำหน้าที่ต่างกัน 2 ค่า คือ ดัชนีคิดเครื่องหมาย (signed index) และดัชนีไม่คิดเครื่องหมาย (unsigned index) ข้อดีของวิธีนี้คือคำนวณง่าย เสียค่าใช้จ่ายน้อย นำไปใช้อธิบายธรรมชาติของข้อสอบทำหน้าที่ต่างกันได้ ส่วนข้อด้อยคือต้องใช้กลุ่มตัวอย่างขนาดใหญ่

กลุ่มวิธีทฤษฎีการตอบสนองข้อสอบ (IRT) เป็นวิธีที่วิเคราะห์ความแตกต่างของฟังก์ชันการตอบสนองข้อสอบ โดยเปรียบเทียบโค้งคุณลักษณะข้อสอบ (item characteristic curves :ICCs) ของกลุ่มผู้สอบตามระดับความสามารถผู้สอบ ถ้าโค้งคุณลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มมีรูปร่างเหมือนกันแสดงว่าข้อสอบข้อนั้นทำหน้าที่ไม่ต่างกัน แต่ถ้าโค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่มมีรูปร่างต่างกันแสดงว่าข้อสอบข้อนั้นทำหน้าที่ต่างกัน กระบวนการวิเคราะห์ใช้สถิติพาราเมตริก มีดัชนีบอกระดับการทำหน้าที่ต่างกันและทดสอบนัยสำคัญทางสถิติ ซึ่งได้รับการพัฒนาและแตกย่อยออกไปหลายวิธีเช่น วิธี General IRTLR วิธี Loglinear IRTLR วิธี IRT-$\hat{b}$และวิธี Lord 's χ^2 จากการศึกษาของนักวิจัยพบว่าวิธีนี้เป็นวิธีที่มีความถูกต้องสูง สามารถตรวจสอบพบจานวนข้อสอบที่ทำหน้าที่ต่างกันมากที่สุด (Holland and Thayer, 1988) ค่าสถิติของข้อสอบไม่เปลี่ยนไปตามกลุ่มตัวอย่าง และใช้การประมาณค่าความสามารถที่แท้จริงของผู้สอบแทนคะแนนที่สังเกตได้ แต่มีข้อจำกัดคือต้องใช้กลุ่มตัวอย่างขนาดใหญ่ ข้อมูลต้องเป็นไปตามข้อตกลงเบื้องต้น กล่าวคือถ้าใช้ข้อมูลจำลองต้องสร้างขึ้นภายใต้ทฤษฎี IRT และมีการคำนวณซับซ้อนหลายรอบ แปลผลยาก ค่าใช้จ่ายสูง (Ryan, 1991 ; Osterlind, 1993 ; Narayanan and Swaminatan,1994 ; สุรศักดิ์ อมรรัตนศักดิ์,2531)

วิธีซิบเทสท์ (SIBTEST) เป็นวิธีที่มีพื้นฐานอยู่บนทฤษฎีการตอบข้อสอบแบบพหุมิติ ใช้คะแนนรวมจากแบบสอบเป็นตัวแทนความสามารถโดยมีข้อตกลงเบื้องต้นว่ามีมิติการวัด 2 มิติ ดังนั้นคะแนนรวมจากแบบสอบจึงมี 2 ส่วนคือคะแนนแบบสอบที่มีความตรง (valid subtest) ซึ่งวัดคุณลักษณะแฝงเป้าหมาย และคะแนนรวมจากแบบสอบที่ศึกษา (studied subtest) ซึ่งวัดคุณลักษณะแฝงแทรกซ้อน มีดัชนีบอกระดับการทำหน้าที่ต่างกันของข้อสอบและการทดสอบนัยสำคัญทางสถิติ วิธีนี้มีประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปได้ใกล้เคียงกับวิธี MH (Shealy and Stout, 1993 ; Narayanan and Swaminatan, 1994)

โดยสรุป วิธีการที่ใช้ตรวจสอบการทำหน้าที่ต่างกันแต่ละวิธีนั้นมีขั้นตอนการวิเคราะห์ จุดเด่น และจุดด้อยแตกต่างกันดังนี้

ตารางที่ 1 เปรียบเทียบวิธีการต่างๆในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ

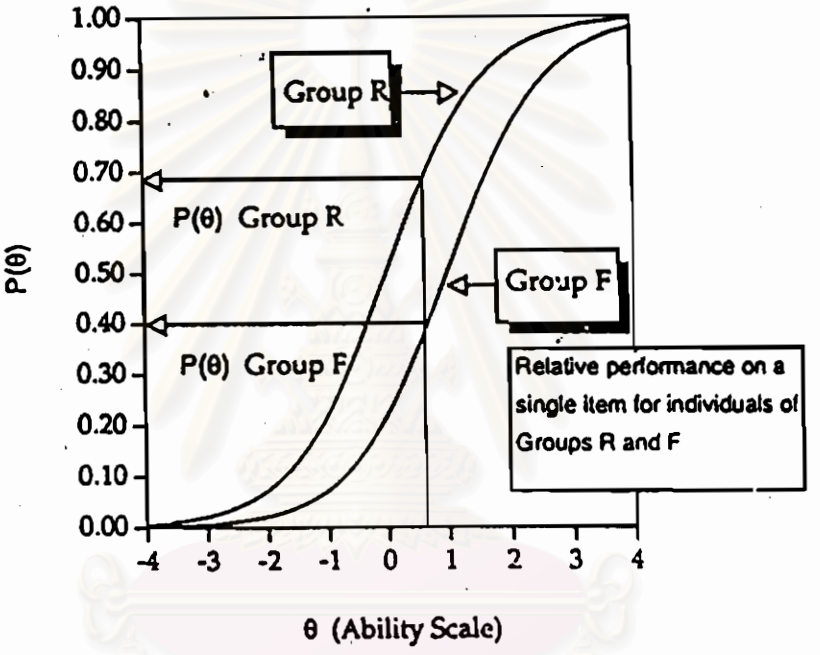
ประเด็น	วิธี IRT	วิธี SIBTEST	วิธี LR	วิธี MH	วิธี STND
1. ข้อตกลงเบื้องต้น	แบบสอบมีมิติเดียวและโค้งลักษณะข้อสอบแสดงฟังก์ชันค่าความสามารถและโอกาสในการตอบข้อสอบถูก	คะแนนรวมของแบบสอบเป็นตัวแทนความสามารถมีมิติการวัด 2 มิติ คือมิติลักษณะแฝงเป้าหมาย (θ)และมิติลักษณะแฝงแทรกซ้อน (η)	คะแนนรวมของแบบสอบเป็นตัวทำนายโอกาสการตอบข้อสอบได้ถูกต้อง	คะแนนรวมของแบบสอบเป็นตัวแทนความสามารถของผู้สอบแต่ละกลุ่ม	คะแนนรวมของแบบสอบเป็นตัวแทนความสามารถของผู้สอบ
2.ทฤษฎีที่เกี่ยวข้อง	IRT	IRT ชนิดพหุมิติ	โมเดลถดถอยโลจิสติก	ไคสแควร์	คะแนนมาตรฐาน

ตารางที่ 1 (ต่อ)

ประเด็น	วิธี IRT	วิธี SIBTEST	วิธี LR	วิธี MH	วิธี STND
3. สิ่งที่ทำการวิเคราะห์	ความแตกต่างของฟังก์ชันการตอบข้อสอบ	ความแตกต่างระหว่างคะแนนเฉลี่ยและสัดส่วนการตอบข้อสอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน	การทำนายโอกาสในการตอบข้อสอบถูกหรือผิด ของผู้สอบสองกลุ่มกับปฏิสัมพันธ์ระหว่างกลุ่ม	ความแตกต่างของสัดส่วนการตอบระหว่างผู้ที่มีความสามารถระดับเดียวกัน	การประมาณค่าความน่าจะเป็นในการตอบข้อสอบถูกของผู้สอบที่มีความสามารถระดับเดียวกัน
4เกณฑ์การทำหน้าที่ต่างกัน	ทดสอบความแตกต่างของพื้นที่ระหว่างโค้งลักษณะข้อสอบด้วยสถิติ Z	ค่าดัชนี β_U และการทดสอบนัยสำคัญค่า Z	การประมาณค่า τ และการทดสอบนัยสำคัญค่า χ^2	ค่าดัชนี α_{MH} และการทดสอบ χ^2_{MH}	ค่าดัชนี $STND_{PDIF}$ และตรวจสอบความคลาดเคลื่อนมาตรฐาน
5.จุดเด่น	ตรวจพบข้อสอบที่ DIF ได้ถูกต้องและการไม่แปรเปลี่ยนค่าพารามิเตอร์	ใช้กลุ่มตัวอย่างน้อย คำนวณง่าย เมื่อการกระจายความสามารถไม่เท่ากันสามารถตรวจสอบ DIF ได้ดี	ตรวจสอบ DIF ได้ทั้งแบบเอกรูปและแบบอเนกรูป	ใช้กลุ่มตัวอย่างน้อย คำนวณง่าย ประหยัดเวลาและค่าใช้จ่าย	คำนวณง่าย ประหยัด อธิบายธรรมชาติและสาเหตุของ DIF ได้ดี
6.จุดด้อย	ใช้กลุ่มตัวอย่างขนาดใหญ่ การคำนวณซับซ้อน แปลผลยาก เสียค่าใช้จ่ายสูง	อัตราความคลาดเคลื่อนชนิดที่ 1 สูงเมื่อคะแนนเฉลี่ยระหว่างกลุ่มต่างกัน	ใช้เวลาและเสียค่าใช้จ่ายในการคิดคำนวณสูง	ตรวจสอบ DIF ได้ไม่ดีเมื่อการกระจายความสามารถระหว่างกลุ่มไม่เท่ากัน	ใช้กลุ่มตัวอย่างขนาดใหญ่ทั้งนี้เพื่อแก้ปัญหาความคลาดเคลื่อนในการสุ่มตัวอย่าง

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ พบว่ามีการระบุลักษณะของข้อสอบที่ทำหน้าที่ต่างกัน ใน 2 ประเภท (Mellenbergh, 1982) คือ

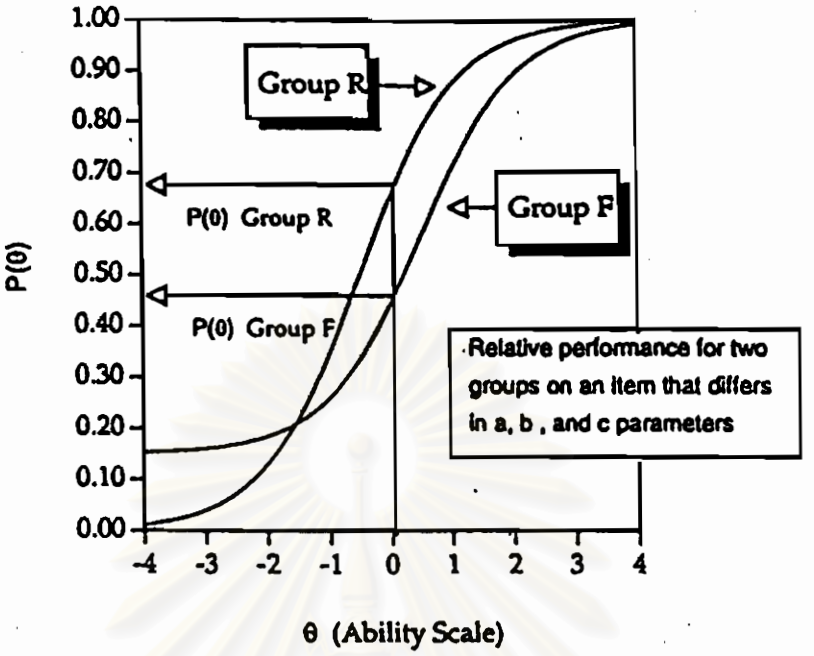
1. ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป (uniform DIF) หมายถึงข้อสอบที่ทำให้ผู้สอบกลุ่มหนึ่งมีโอกาสในการตอบข้อสอบถูกมากกว่าผู้สอบอีกกลุ่มหนึ่งสม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่มจะพบว่าไม่มีปฏิสัมพันธ์ระหว่างโค้งคุณลักษณะข้อสอบในทุกระดับความสามารถ



ภาพที่ 1 ข้อสอบทำหน้าที่ต่างกันแบบเอกรูป

2. ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป (nonuniform DIF) หมายถึงข้อสอบที่ทำให้โอกาสในการตอบข้อสอบถูกของผู้สอบระหว่างกลุ่มไม่สม่ำเสมอกันในทุกระดับความสามารถ เมื่อพิจารณาโค้งคุณลักษณะข้อสอบของผู้สอบ 2 กลุ่มพบว่ามีปฏิสัมพันธ์ร่วมกันระหว่างโค้งคุณลักษณะ เช่น ที่ระดับความสามารถหนึ่ง กลุ่มผู้สอบกลุ่ม R มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม F แต่ที่ระดับความสามารถอีกระดับหนึ่งกลุ่มผู้สอบกลุ่ม F มีโอกาสในการตอบข้อสอบถูกมากกว่ากลุ่มผู้สอบกลุ่ม R

หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย



ภาพที่ 2 ข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป

ตอนที่ 2 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีทฤษฎีการตอบสนองข้อสอบ

วิธีทฤษฎีการตอบสนองข้อสอบเป็นวิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยพิจารณาเปรียบเทียบฟังก์ชันการตอบข้อสอบ (IRFs) ระหว่างกลุ่มผู้สอบที่มีต่อแบบสอบชุดเดียวกันถ้าฟังก์ชันการตอบข้อสอบระหว่างกลุ่มผู้สอบต่างกัน แสดงว่าข้อสอบทำหน้าที่ต่างกัน (Kim et al., 1994) หรือเป็นการเปรียบเทียบความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบ (ICCs) ระหว่างกลุ่มผู้สอบย่อยสองกลุ่มและใช้พื้นที่ระหว่างโค้งคุณลักษณะข้อสอบเป็นดัชนีแสดงข้อสอบทำหน้าที่ต่างกัน (Osterlind, 1992) โดยที่ข้อสอบทำหน้าที่ต่างกันแบบเอกรูปเกิดขึ้นเมื่อมีความแตกต่างระหว่างฟังก์ชันการตอบข้อสอบ (IRF_S) ระหว่างกลุ่มผู้สอบคงที่ในทุกระดับความสามารถ ส่วนข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปเกิดขึ้นเมื่อมีความแตกต่างระหว่างฟังก์ชันการตอบข้อสอบระหว่างกลุ่มผู้สอบไม่คงที่ในทุกระดับความสามารถ (Millsap and Everson, 1993) ในการวิเคราะห์ใช้ทฤษฎีการตอบสนองข้อสอบแบบ 2 หรือ 3 พารามิเตอร์ ซึ่งมีข้อตกลงเบื้องต้นที่สำคัญคือ คะแนนผลการตอบข้อสอบต้องมาจากแบบสอบที่มุ่งวัดลักษณะที่สำคัญเพียงลักษณะเดียว (unidimensional) และโมเดลโลจิสติกแบบ 2 หรือ 3 พารามิเตอร์ สามารถใช้แทนข้อมูลผลการตอบข้อสอบได้อย่างพอเพียง

ในการวัดพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบเพื่อพิจารณาข้อสอบทำหน้าที่ต่างกันมี 2 ลักษณะคือ การวัดช่วงเปิด และการวัดช่วงปิด (open and closed interval) การวัดช่วงเปิด จะเป็นการวัดในช่วงความสามารถทั้งหมดระหว่างโค้งทั้งสอง ซึ่งจะทำให้ได้พื้นที่แน่นอนระหว่างโค้งสองโค้ง ส่วนการวัดช่วงปิดจะกำหนดขอบเขตการวัดให้อยู่ในช่วงความสามารถที่กำหนด มีดัชนีที่ใช้บอกระดับการทำหน้าที่ต่างกัน ได้แก่ พื้นที่ขนิดไม่มีเครื่องหมาย (unsigned areas) เป็นค่าสัมบูรณ์ของพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบ และพื้นที่ขนิดมีเครื่องหมาย (signed areas) เหมือนกับพื้นที่ขนิดไม่มีเครื่องหมาย แต่จะมีเครื่องหมายแสดงให้ทราบว่ากลุ่มใดได้ประโยชน์ กลุ่มใดเสียประโยชน์ ถ้าโค้งการตอบข้อสอบตัดกันแสดงว่ามีการทำหน้าที่ต่างกันสำหรับผู้สอบที่มีความสามารถต่างกัน

จากการศึกษาของ Kim และคณะ (1994) พบว่าแบบจำลองของทฤษฎีการตอบสนองข้อสอบแบบ 2 พารามิเตอร์มีความคลาดเคลื่อนประเภทที่ 1 ต่ำกว่าแบบจำลอง 3 พารามิเตอร์และการใช้โมเดล 2 พารามิเตอร์และกลุ่มตัวอย่าง 1000 คน เป็นเงื่อนไขที่ให้ผลการประมาณค่าดีที่สุด ซึ่งวิธี IRT แบบ 2 พารามิเตอร์ มีสูตรในการ คำนวณดังนี้ (Raju, 1988 cited in Cohen et al., 1991)

$$P_i(\theta) = \left[1 + \exp\{-1.7a_i(\theta - b_i)\}\right]^{-1}$$

$$= \left[1 + \exp\{-L_i(\theta)\}\right]^{-1}$$

$$= \frac{\exp\{-L_i(\theta)\}}{1 + \exp\{-L_i(\theta)\}}$$

โดยที่ $L_i(\theta) = 1.7a_i(\theta - b_i)$

พื้นที่ (S_i) ภายใต้โค้งคุณลักษณะข้อสอบสำหรับข้อสอบข้อที่ i ระหว่างจุดสองจุดซึ่งเป็นช่วงที่สนใจ (θ_1, θ_2) สามารถเขียนได้ดังนี้

$$S_i(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} P_i(\theta)d\theta$$

$$= (1.7a_i)^{-1} \ln\left[1 + \exp\{L_i(\theta)\}\right] \Big|_{\theta_1}^{\theta_2}$$

$$= (1.7a_i)^{-1} \left[\ln\left[1 + \exp\{L_i(\theta_2)\}\right] - \ln\left[1 + \exp\{L_i(\theta_1)\}\right]\right]$$

โมเดล 2 พารามิเตอร์มีค่าพารามิเตอร์ข้อสอบอยู่สองค่าสำหรับข้อสอบข้อที่ i สำหรับกลุ่มอ้างอิง (a_{iR}, b_{iR}) และกลุ่มเปรียบเทียบ (a_{iF}, b_{iF}) สำหรับพื้นที่แบบมีเครื่องหมาย (SA) และพื้นที่แบบไม่มีเครื่องหมาย (UA) ระหว่างโค้งคุณลักษณะข้อสอบในช่วง θ_1, θ_2 สามารถคำนวณได้ดังนี้

$$SA = \int_{\theta_1}^{\theta_2} \{P_{iR}(\theta) - P_{iF}(\theta)\} d\theta$$

$$= \int_{\theta_1}^{\theta_2} P_{iR}(\theta) d\theta - \int_{\theta_1}^{\theta_2} P_{iF}(\theta) d\theta$$

$$= S_{iR}(\theta_1, \theta_2) - S_{iF}(\theta_1, \theta_2)$$

$$UA = \int_{\theta_1}^{\theta_2} | P_{iR}(\theta) - P_{iF}(\theta) | d\theta$$

เมื่อ θ_1, θ_2 คือระดับความสามารถที่ต่ำกว่าและสูงกว่าตามลำดับ

$P_{iR}(\theta), P_{iF}(\theta)$ คือความน่าจะเป็นในการตอบข้อสอบข้อที่ i ได้ถูกต้องของผู้สอบที่ระดับความสามารถนั้นของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบตามลำดับ

S_{iR}, S_{iF} คือพื้นที่ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบสำหรับข้อสอบข้อที่ i

ถ้าโค้งคุณลักษณะข้อสอบทั้งสองโค้งไม่ตัดกัน ($a_{iR} = a_{iF}$) สามารถคำนวณหาพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบข้อที่ i ซึ่งเป็นพื้นที่ไม่คิดเครื่องหมายได้โดยพื้นที่ที่คำนวณได้จะเท่ากับพื้นที่แบบมีเครื่องหมาย

$$UA = | S_{iR}(\theta_1, \theta_2) - S_{iF}(\theta_1, \theta_2) |$$

$$= | SA |$$

ถ้าโค้งคุณลักษณะข้อสอบทั้งสองโค้งตัดกัน ($a_{iR} \neq a_{iF}$) ที่ θx

$$\theta x = \frac{(a_{iF} b_{iF} - a_{iR} b_{iR})}{(a_{iF} - a_{iR})}$$

สามารถคำนวณหาพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบข้อที่ i ซึ่งเป็นพื้นที่ไม่คิดเครื่องหมายได้ดังนี้

$$UA = | S_{iR}(\theta_1, \theta x) - S_{iF}(\theta_1, \theta x) | + | S_{iR}(\theta_x, \theta_2) - S_{iF}(\theta x, \theta_2) |$$

$$= \left| \int_{\theta_1}^{\theta x} \{P_{iR}(\theta) - P_{iF}(\theta)\} d\theta \right| + \left| \int_{\theta_x}^{\theta_2} \{P_{iR}(\theta) - P_{iF}(\theta)\} d\theta \right|$$

$$= \left| \int_{\theta_1}^{\theta x} P_{iR}(\theta) d\theta - \int_{\theta_1}^{\theta x} P_{iF}(\theta) d\theta \right| + \left| \int_{\theta_x}^{\theta_2} P_{iR}(\theta) d\theta - \int_{\theta_x}^{\theta_2} P_{iF}(\theta) \} d\theta \right|$$

ตอนที่ 3 วิธีตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล (MH)

วิธี MH เป็นวิธีที่พัฒนามาจากวิธีไคสแควร์แบบดั้งเดิม (traditional χ^2 approach) ซึ่งเสนอโดย Mantel และ Haenzel ในปี ค.ศ. 1959 เพื่อนำไปใช้นงานวิจัยทางการแพทย์ แต่ผู้ที่เริ่มนำมาใช้เพื่อตรวจสอบการทำหน้าที่ต่างกันของแบบสอบคือ Holland (1985) Holland และ Thayer (1988) หลังจากนั้นวิธี MH ถูกนำมาใช้อย่างกว้างขวางเพราะเป็นวิธีที่คำนวณง่าย ประหยัด ใช้กลุ่มตัวอย่างน้อย และการแปลผลไม่ยุ่งยาก

หลักการตรวจสอบการทำหน้าที่ต่างกันของวิธี MH เป็นการเปรียบเทียบผลการสอบของผู้สอบ 2 กลุ่ม คือกลุ่มอ้างอิง (reference group) และกลุ่มเปรียบเทียบ (focal group) กลุ่มอ้างอิงคือกลุ่มที่คาดว่าจะได้ประโยชน์ในการตอบข้อสอบ ส่วนกลุ่มเปรียบเทียบเป็นกลุ่มที่คาดว่าจะเสียประโยชน์จากการตอบข้อสอบในกรณีข้อสอบทำหน้าที่ต่างกัน และในการตรวจสอบครั้งหนึ่งๆจะเรียกข้อสอบที่ถูกตรวจสอบการทำหน้าที่ต่างกันว่า ข้อสอบที่ศึกษา (siudied item) ซึ่งจะมีการตรวจสอบกลุ่มผู้สอบทุกๆระดับคะแนนรวมจากแบบสอบ ข้อสอบใดที่ผู้สอบสองกลุ่มทำได้เท่ากันถือว่าเป็นข้อสอบทำหน้าที่ไม่ต่างกัน (no DIF) โดยจะพิจารณาจากผลการตอบข้อสอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่มีคะแนนรวมเท่ากันแล้วแสดงความถี่ของผู้สอบทั้งสองกลุ่มที่ตอบข้อสอบนั้นถูก (1) หรือผิด (0) ลงในตาราง 2 ทาง

ตารางที่ 2 ความถี่ของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

กลุ่ม	คะแนนจากข้อสอบที่ต้องการตรวจสอบ DIF		
	1	0	รวม
อ้างอิง	A_j	B_j	N_{Rj}
เปรียบเทียบ	C_j	D_j	N_{fj}
รวม	m_{1j}	m_{0j}	N_j

เมื่อ N_j เป็นความถี่ของผู้สอบทั้งหมดที่ได้ระดับคะแนน j

N_{Rj} , N_{fj} เป็นความถี่ของผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ได้ระดับคะแนน j

m_{1j} , m_{0j} เป็นความถี่ของผู้สอบที่ตอบข้อสอบถูกและผิดที่ระดับคะแนน j

A_j , B_j , C_j , D_j เป็นความถี่ของผู้สอบที่ตอบถูก(1) และตอบผิด(0) ของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

ตารางที่ 3 สัดส่วนการตอบข้อสอบของกลุ่มอ้างอิงและกลุ่มเปรียบเทียบที่ระดับคะแนน j

กลุ่ม	คะแนนจากข้อสอบที่ต้องการตรวจสอบ DIF		
	1	0	รวม
อ้างอิง	P_{Rj}	Q_{Rj}	1
เปรียบเทียบ	P_{Fj}	Q_{Fj}	1

เมื่อ P_{Rj} , Q_{Rj} คือสัดส่วนการตอบข้อสอบของกลุ่มอ้างอิงที่ตอบข้อสอบถูกและผิดตามลำดับ

P_{Fj} , Q_{Fj} คือสัดส่วนการตอบข้อสอบของกลุ่มเปรียบเทียบที่ตอบข้อสอบถูกและผิดตามลำดับ

ดังนั้นจึงมีสมมติฐานศูนย์ดังนี้ $H_0 : P_{Rj} = P_{Fj}$ ทุกค่าของ j

หรือ $H_0 : \frac{A_j D_j}{N_j} = \frac{B_j C_j}{N_j}$ ทุกค่าของ j

ขั้นตอนการวิเคราะห์ดำเนินการตามขั้นตอนดังนี้

1. คำนวณความน่าจะเป็นในรูปสัดส่วนการตอบข้อสอบถูกและผิดระหว่างกลุ่ม ในแต่ละข้อในช่วงคะแนน j โดยใช้สูตร

$$\alpha_{MH} = \frac{\sum A_j D_j / N_j}{\sum B_j C_j / N_j}$$

ค่า α_{MH} มีค่าระหว่าง 0 ถึง α ถ้าค่า α_{MH} ที่คำนวณได้มีค่าเท่ากับ 1 แสดงว่าข้อสอบทำหน้าที่ไม่แตกต่างกัน ถ้าค่า $\alpha_{MH} > 1$ แสดงว่าข้อสอบเข้าข้างกลุ่มอ้างอิง

2. ทดสอบนัยสำคัญของค่าไคสแควร์ เพื่อทดสอบค่าที่คำนวณได้จะมีความแตกต่างจาก 1 อย่างมีนัยสำคัญหรือไม่ที่ระดับชั้นความมีอิสระเท่ากับ 1

$$\chi^2_{MH} = \frac{\{|\sum A_j - \sum E(A_j)| - 0.5\}^2}{\sum Var(A_j)}$$

เมื่อ $E(A_j) = (N_{Rj})(m_{1j}) / N_j$

$$Var(A_j) = \frac{N_{Rj} N_{Fj} m_{1j} m_{0j}}{N_j^2 (N_j - 1)}$$

Holland และ Thayer ได้เสนอให้แปลงค่า α_{MH} ให้เป็นค่าเดลต้า (Δ_{MH}) หรือ MH_{DIF} ดังนี้

$$MH_{DIF} = -2.35 \ln (\alpha_{MH})$$

ค่า MH_{DIF} มีค่าระหว่าง -2.6 ถึง 2.6 ถ้าค่า $MH_{DIF} = 0$ แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน ถ้า MH_{DIF} มีค่าเป็นบวกแสดงว่าเข้าข้างกลุ่มเปรียบเทียบ และถ้า MH_{DIF} มีค่าเป็นลบแสดงว่าเข้าข้างกลุ่มอ้างอิง

ตอนที่ 4 วิธีการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีถดถอยโลจิสติก (LR)

วิธีถดถอยโลจิสติก (LR) เป็นวิธีที่พัฒนามาจากวิธี Loglinear ของ Mellenberg เสนอโดย Swaminatan และ Rogers ในปี ค.ศ. 1990 วิธีนี้อยู่บนพื้นฐานของโมเดลถดถอยโลจิสติก ซึ่งเป็นโมเดลที่มีพื้นฐานเป็นแบบจำลองที่สามารถเพิ่มตัวแปรความสามารถเข้าไปในแบบจำลองได้ และเข้าใจธรรมชาติของการทำหน้าที่ต่างกันของแบบสอบได้ดี สามารถตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปและแบบอเนกรูปได้

โมเดลถดถอยโลจิสติกเป็นโมเดลที่มีความยืดหยุ่นและใช้ได้ง่าย และในการวิเคราะห์สมการถดถอยโลจิสติกหรือเรียกว่าการวิเคราะห์โลจิท (Logit Anlaysis) เป็นการวิเคราะห์สมการทำนายเมื่อต้องการศึกษาผลของตัวแปรทำนาย (predictor variable) ที่มีตัวแปรเกณฑ์เป็นทวิวิภาค (dichotomous variable) โดยใช้ฟังก์ชันโลจิสติก (logistic function) ในการแสดงความสัมพันธ์ระหว่างค่าของตัวแปรทำนายกับค่าความน่าจะเป็นของการเกิดเหตุการณ์ตามตัวแปรเกณฑ์ (ศิริชัย กาญจนวาสี, 2541)

เมื่อนำมาวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบ เป็นการวิเคราะห์โดยใช้กลุ่มผู้สอบ (กลุ่ม 1 หรือ กลุ่ม 2) ความสามารถผู้สอบ (θ) และปฏิสัมพันธ์ระหว่างกลุ่มผู้สอบกับความสามารถผู้สอบในการทำนายโอกาสในการตอบข้อสอบถูกหรือผิด ดังนี้

$$P(U = 1/\theta) = \frac{e^{(\beta_0 + \beta_1\theta)}}{1 + e^{(\beta_0 + \beta_1\theta)}}$$

เมื่อ U เป็นคำตอบของข้อสอบ

θ เป็นความสามารถของผู้สอบแต่ละคน

β_0 เป็นพารามิเตอร์จุดตัด (intercept parameter)

β_1 เป็นค่าพารามิเตอร์ความชัน (slope parameter)

ในการวิเคราะห์การทำหน้าที่ต่างกันของข้อสอบโดยแบ่งสมการสำหรับผู้สอบออกเป็น 2 กลุ่มดังนี้

$$P(U_{ij} = 1/\theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}}{1 + e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}} \quad , i = 1,.....n_j \quad , j = 1,2$$

เมื่อ U_{ij} เป็นคำตอบข้อสอบของบุคคลที่ i ในกลุ่ม j

θ_{ij} เป็นความสามารถของผู้สอบที่ i ในกลุ่ม j

β_{0j} เป็นพารามิเตอร์จุดตัด (intercept parameter)

β_{1j} เป็นค่าพารามิเตอร์ความชันของผู้สอบในกลุ่ม j (slope parameter)

ถ้า $\beta_{01} = \beta_{02}$ และ $\beta_{11} = \beta_{12}$ เป็นลักษณะโค้งถดถอยโลจิสติกสำหรับผู้สอบสองกลุ่มเหมือนกัน แสดงว่าข้อสอบทำหน้าที่ไม่ต่างกัน

$\beta_{01} \neq \beta_{02}$ แต่ $\beta_{11} = \beta_{12}$ เป็นลักษณะโค้งถดถอยโลจิสติกสำหรับผู้สอบสองกลุ่มเท่าเทียมกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป

$\beta_{01} = \beta_{02}$ แต่ $\beta_{11} \neq \beta_{12}$ เป็นลักษณะโค้งถดถอยโลจิสติกสำหรับผู้สอบสองกลุ่มไม่เท่าเทียมกัน แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป

จากสมการดังกล่าวสามารถเขียนเป็นสมการใหม่ ได้เป็น

$$P(U = 1/\theta) = \frac{e^{z}}{1 + e^{z}}$$

เมื่อ $Z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g)$

g แทนกลุ่มผู้สอบ (g = 1 ผู้สอบเป็นสมาชิกกลุ่ม 1 ,g =2 ผู้สอบเป็นสมาชิกกลุ่ม 2)

θg เป็นผลของตัวแปรอิสระ 2 ตัว คือ θ กับ g

τ_2 เป็นความแตกต่างของกลุ่มในการทำข้อสอบ

$\tau_2 = \beta_{01} - \beta_{02}$

τ_3 เป็นปฏิสัมพันธ์ระหว่างกลุ่มกับความสามารถผู้สอบ

$\tau_3 = \beta_{11} - \beta_{12}$

ถ้า $\tau_2 \neq 0$ และ $\tau_3 = 0$ แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบเอกรูป

$\tau_3 \neq 0$ ($\tau_2 = 0$ หรือไม่ก็ได้) แสดงว่าข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป

การประมาณค่าพารามิเตอร์ใช้วิธีประมาณค่าด้วยวิธีแมกซิมัมไลลิฮูด(maximum likelihood) ส่วนการทดสอบนัยสำคัญทางสถิติใช้วิธีตรวจสอบนัยสำคัญของค่าไคสแควร์ โดยมีสมมติฐานศูนย์ดังนี้

$H_0 : \tau_2 = 0$ และ $H_0 : \tau_3 = 0$ ทดสอบในแต่ละครั้ง

ถ้าสมมติฐานสูญทั้งสองตรวจสอบพร้อมกัน จะมีสมมติฐานใหม่ดังนี้

$H_0 : C\tau = 0$

เมื่อ C เป็นเมตริก 2×4

และทดสอบนัยสำคัญด้วยค่าไคสแควร์ที่ขั้นความเป็นอิสระเท่ากับ 2

$$\chi^2 = \hat{\tau}' C' (C \sum C')^{-1} C \hat{\tau}'$$

ตอนที่ 5 งานวิจัยที่เกี่ยวข้อง

Swaminathan และ Rogers (1990) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป โดยใช้การจำลองข้อมูล ขนาดกลุ่มตัวอย่าง 2 ขนาดคือ 250 และ 500 คน ความยาวแบบสอบ 3 ขนาด คือ 40 60 และ 80 ข้อ จำลองข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปโดยให้ค่าอำนาจจำแนกของข้อสอบสองกลุ่มเท่ากัน ส่วนค่าความยากผันแปรไปเพื่อให้ได้ข้อสอบที่ทำหน้าที่ต่างกันตามต้องการ กล่าวคือในแต่ละความยาวแบบสอบจะมีข้อสอบที่ทำหน้าที่ต่างกัน 20% ของความยาวแบบสอบ ส่วนข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปนั้นกำหนดให้ค่าความยากของสองกลุ่มเท่ากันแต่ค่าอำนาจจำแนกของข้อสอบแปรเปลี่ยนไป

ผลการศึกษาพบว่าวิธีถดถอยโลจิสติกและวิธีแมนเทล-แฮนส์เซลให้ผลใกล้เคียงกันในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูป กล่าวคือมีการตรวจค้นได้ถูกต้องร้อยละ 70 กรณีใช้กลุ่มตัวอย่าง 250 คน และตรวจค้นได้ร้อยละ 100 ในกลุ่มตัวอย่าง 500 คน ในทุกความยาวแบบสอบ สำหรับการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูปพบว่าวิธีแมนเทล-แฮนส์เซลตรวจค้นได้เล็กน้อย ส่วนวิธีถดถอยโลจิสติกตรวจค้นได้ถูกต้องร้อยละ 50 ในกรณีกลุ่มตัวอย่างน้อยและข้อสอบสั้น และถูกต้องร้อยละ 75 ในแบบสอบยาวและตัวอย่างขนาดใหญ่

สำหรับการตรวจค้นอัตราความคลาดเคลื่อนชนิดที่ 1 พบว่าวิธีแมนเทล-แฮนส์เซล ตรวจค้นผิดพลาดร้อยละ 1 ส่วนวิธีถดถอยโลจิสติกตรวจค้นผิดพลาดร้อยละ 1-6

Mazor และคณะ (1992) ศึกษาปัจจัยขนาดกลุ่มตัวอย่างที่มีผลต่อการตรวจสอบการทำหน้าที่ต่างกันของวิธีแมนเทล-แฮนส์เซล โดยศึกษาจากข้อมูลจำลอง กลุ่มตัวอย่างที่ใช้มี 5 ขนาด คือ 100 200 500 1000 และ 2000 คน ความยาวแบบสอบ 75 ข้อ ตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันด้วยการทดสอบ MH-χ^2 ผลการศึกษาพบว่า เมื่อใช้กลุ่มตัวอย่างขนาดใหญ่ ทำให้ตรวจพบข้อสอบที่ทำหน้าต่างกันถูกต้องมากกว่าการใช้กลุ่มผู้สอบขนาดเล็ก กล่าวคือ เมื่อขนาดกลุ่มตัวอย่างเท่ากับ 2000 คน ตรวจค้นข้อสอบได้ถูกต้องร้อยละ 70 ถึงร้อยละ 75 แต่เมื่อใช้กลุ่มตัวอย่าง 500 คนหรือน้อยกว่าสามารถตรวจค้นข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องน้อยกว่าร้อยละ 50 และกล่าวว่าการใช้กลุ่มผู้สอบขนาด 200 ถึง 1000 คนต่อกลุ่ม มีความเพียงพอต่อการตรวจสอบการทำหน้าที่ต่างกัน นอกจากนี้ข้อสอบที่ไม่สามารถตรวจค้นการทำหน้าที่ต่างกันได้เนื่องจากข้อสอบมีความยากมากหรือมีความยากต่างกันเพียงเล็กน้อยระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบ และเป็นข้อสอบที่มีค่าอำนาจจำแนกต่ำ

Rogers และ Swaminathan (1993) ได้เปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ โดยใช้ข้อมูลจำลอง ศึกษาเกี่ยวกับการกระจายของสถิติทดสอบและประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปและแบบอเนกรูป การศึกษาด้านการกระจายของสถิติทดสอบ ศึกษาปัจจัยที่แปรเปลี่ยนคือ ขนาดกลุ่มตัวอย่าง (250 และ 500 คน) ความเหมาะสมของข้อมูลกับโมเดล ค่าความยากและค่าอำนาจจำแนกของข้อสอบ ใช้ความยาวแบบสอบ 40 ข้อ ส่วนการศึกษาด้านประสิทธิภาพการตรวจสอบได้ศึกษาปัจจัยที่แปรเปลี่ยนคือ ขนาดกลุ่มตัวอย่าง (250 และ 500 คน ความเหมาะสมของข้อมูลกับโมเดล ความยาวแบบสอบ (40 และ 80 ข้อ) การกระจายของคะแนนสอบ (การกระจายของคะแนนสอบแบบโค้งปกติและแบบเบ้ลบ) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกัน (0% และ 15%) ค่าความยาก ค่าอำนาจจำแนก และขนาดของข้อสอบที่ทำหน้าที่ต่างกันหรือพื้นที่ระหว่างโค้งคุณลักษณะข้อสอบระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบ

ผลการศึกษาด้านการกระจายของสถิติทดสอบพบว่าการกระจายของสถิติทดสอบของทั้งสองวิธีอยู่ในระดับที่น่าพอใจ ยกเว้นการกระจายของสถิติของวิธีถดถอยโลจิสติกไม่เป็นตามที่คาดไว้ในกรณีที่ข้อสอบยากมากและค่าอำนาจจำแนกสูง ด้านประสิทธิภาพการตรวจสอบพบว่าทั้ง

สองวิธีมีประสิทธิภาพในการตรวจสอบเท่ากันในการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูป แต่วิธีถดถอยโลจิสติกมีประสิทธิภาพมากกว่าในการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป ขนาดกลุ่มตัวอย่างเป็นปัจจัยที่มีผลกระทบอย่างมากในการตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป คือเมื่อเพิ่มขนาดกลุ่มตัวอย่างอัตราการตรวจสอบจะเพิ่มขึ้น ส่วนขนาดของแบบสอบและการกระจายของคะแนนไม่มีผลต่ออัตราการตรวจสอบ วิธีถดถอยโลจิสติกตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ดีในกรณีข้อสอบยากง่ายปานกลางและค่าอำนาจจำแนกสูง ส่วนวิธีแมนเทล-แฮนส์เซลตรวจสอบข้อสอบยากง่ายปานกลางได้น้อยมาก แต่วิธีนี้สามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ดีในกรณีข้อสอบง่ายง่ายและข้อสอบยาก

Clauser และคณะ(1993) ศึกษาเกณฑ์การจับคู่กลุ่มผู้สอบที่มีความบริสุทธิ์ (purification of the matching criterion) ในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบด้วยวิธีแมนเทล-แฮนส์เซล ระหว่างเทคนิค 1 ขั้นตอน (one step procedure) และเทคนิคแบบ 2 ขั้นตอน (two step procedure) ศึกษาจากข้อมูลจำลองใช้กลุ่มผู้สอบกลุ่มอ้างอิงและกลุ่มเปรียบเทียบกลุ่มละ 1000 คน ความยาวแบบสอบ 3 ขนาด ได้แก่ 20 ข้อ 40 ข้อ และ 80 ข้อ และสร้างข้อสอบทำหน้าที่ต่างกันลงในแบบสอบแต่ละขนาดจำนวน 0% 3% 8% และ 20%

ผลการศึกษาพบว่าผลการตรวจพบข้อสอบทำหน้าที่ต่างกันด้วยวิธีแมนเทล-แฮนส์เซลด้วยเทคนิค 2 ขั้นตอนได้เท่ากับหรือสูงกว่าแบบขั้นตอนเดียวในทุกเงื่อนไขการทดสอบ โดยไม่ทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มสูงขึ้น

Clauser และคณะ(1994) ศึกษาผลจากความกว้างของชั้นคะแนนที่มีต่ออัตราการตรวจสอบและอัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีแมนเทล-แฮนส์เซล ศึกษาจากข้อมูลจำลองความยาวแบบสอบ 80 ข้อ ประกอบด้วยข้อสอบทำหน้าที่ไม่ต่างกันจำนวน 70 ข้อ และข้อสอบทำหน้าที่ต่างกันจำนวน 10 ข้อ ขนาดกลุ่มตัวอย่าง 4 ขนาด คือ 100 200 500 1000 และ 2000 คน และความกว้างของจำนวนชั้นคะแนนแบ่งออกเป็น 2 5 10 20 และ 80

ผลการศึกษาพบว่าในกรณีกลุ่มผู้สอบสองกลุ่มมีการกระจายความสามารถคล้ายกัน การลดจำนวนชั้นคะแนนที่ใช้เป็นเกณฑ์จับคู่กลุ่มผู้สอบไม่มีผลต่ออัตราการตรวจสอบ ในขณะที่อัตราความคลาดเคลื่อนประเภทที่ 1 เพิ่มขึ้นเล็กน้อยหรือไม่เพิ่มขึ้นเลย สำหรับกรณีกลุ่มผู้สอบสองกลุ่มมีการกระจายความสามารถต่างกัน การลดจำนวนชั้นคะแนนที่ใช้เป็นเกณฑ์จับคู่กลุ่มผู้สอบทำให้อัตราการตรวจสอบสูงขึ้นเล็กน้อยแต่อัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้นมาก

Clauser และคณะจึงเสนอแนะว่า การใช้วิธีแมนเทล-แฮนส์เซลตรวจสอบการทำหน้าที่ต่างกันของแบบสอบ ควรใช้จำนวนชั้นคะแนนที่เป็นไปได้มากที่สุดจากคะแนนรวมของผู้สอบ (total score) เป็นเกณฑ์ในการจับคู่กลุ่มผู้สอบ

Mazor และคณะ (1994) ศึกษาการใช้วิธีแมนเทล-แฮนส์เซลตรวจสอบการทำหน้าที่ต่างกันแบบอเนกรูป โดยศึกษาจากข้อมูลจำลอง ใช้กลุ่มตัวอย่างกลุ่มละ 1000 คน ใช้แบบสอบ 25 ฉบับ แต่ละฉบับมี 75 ข้อ ในแต่ละฉบับมีข้อสอบที่ทำหน้าที่ไม่ต่างกัน 59 ข้อ และข้อสอบที่ทำหน้าที่ต่างกัน 16 ข้อ ซึ่งข้อสอบที่ทำหน้าที่ต่างกันจะแปรเปลี่ยนค่าพารามิเตอร์ดังนี้ ค่าอำนาจจำแนก 4 ระดับ (0.25, 0.60, 0.90, 1.25) ความแตกต่างของค่าอำนาจจำแนกระหว่างกลุ่ม 5 ระดับ (0, 0.25, 0.50, 0.75, 1.0) ค่าความยากของข้อสอบกลุ่มอ้างอิง 5 ระดับ (-1.5, -1.0, 0, 1.0, 1.5) ความแตกต่างของค่าความยากระหว่างกลุ่ม 4 ระดับ (0, 0.30, 0.60, 1.0) ค่าการเดากำหนดให้มีค่าเท่ากับ 0.2 และการกระจายความสามารถระหว่างกลุ่มผู้สอบสองแบบคือ การกระจายความสามารถเท่ากันและไม่เท่ากัน

ในการวิเคราะห์ใช้การวิเคราะห์ 3 ครั้ง คือ ครั้งที่ 1 ใช้คะแนนรวมของแต่ละคนเป็นเกณฑ์ในการจับคู่เปรียบเทียบระหว่างกลุ่ม จากนั้นใช้ค่าเฉลี่ยจากคะแนนของกลุ่มตัวอย่างทุกคนเป็นเกณฑ์ในการแบ่งเป็นกลุ่มสูงกลุ่มต่ำ กลุ่มผู้สอบที่ได้คะแนนต่ำกว่าค่าเฉลี่ยจัดให้เป็นกลุ่มที่มีความสามารถต่ำ และกลุ่มที่ได้คะแนนสูงกว่าค่าเฉลี่ยจัดให้เป็นกลุ่มที่มีความสามารถสูง แล้วจึงวิเคราะห์ครั้งที่ 2 คือวิเคราะห์ข้อสอบในกลุ่มผู้สอบที่มีความสามารถสูง ครั้งที่ 3 วิเคราะห์ข้อสอบในกลุ่มผู้สอบที่มีความสามารถต่ำ

ผลการศึกษาพบว่าการวิเคราะห์โดยแบ่งกลุ่มตัวอย่างเป็นกลุ่มสูงกลุ่มต่ำ จะทำให้อัตราการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปสูงกว่าการวิเคราะห์ครั้งที่ 1 (รวมผู้สอบกลุ่มสูงและกลุ่มต่ำไว้ด้วยกัน) อีกทั้งไม่ทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้น และพบว่าเมื่อมีความแตกต่างของค่าอำนาจจำแนกและค่าความยากระหว่างกลุ่มเพิ่มขึ้น จะทำให้อัตราการตรวจสอบพบข้อสอบที่ทำหน้าที่ต่างกันได้มากขึ้น

Uttaro และ Millsap (1994) ศึกษาปัจจัยที่มีอิทธิพลต่อวิธีแมนเทล-แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบ ศึกษาจากข้อมูลจำลอง ปัจจัยที่ศึกษาคือ ความยาวแบบสอบ 20 และ 40 ข้อ ภายใต้เงื่อนไขข้อสอบทำหน้าที่ต่างกันซึ่งเป็นผลเนื่องมาจากฟังก์ชันการตอบข้อสอบของกลุ่มเปรียบเทียบแปรเปลี่ยนไป จึงกำหนดฟังก์ชันการตอบของกลุ่มอ้างอิงคงที่ (a = 1.0 , b = 0.0 , c = 0.2) ส่วนกลุ่มเปรียบเทียบกำหนดค่าอำนาจจำแนกต่างกัน 3 ระดับ ค่าความยากต่างกัน 3 ระดับ และค่าการเดา 2 ระดับ สำหรับเงื่อนไขข้อสอบทำหน้าที่ไม่ต่างกันซึ่งเป็นผลเนื่องมาจากฟังก์ชันการตอบข้อสอบของผู้สอบสองกลุ่มเหมือนกัน จึงกำหนดค่าพารามิเตอร์ของข้อสอบเท่ากันทั้งสองกลุ่ม ขนาดกลุ่มตัวอย่างที่ศึกษากลุ่มละ 500 คน ในทุกเงื่อนไข

ผลการศึกษาพบว่าภายใต้เงื่อนไขข้อสอบทำหน้าที่ไม่ต่างกันความยาวแบบสอบ การกระจายความสามารถ ค่าอำนาจจำแนกและค่าการเดามีผลกระทบต่ออัตราความคลาดเคลื่อนชนิดที่ 1 และการประมาณค่า α_{MH} โดยข้อสอบ 20 ข้อมีผลทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 สูง ส่วนข้อสอบ 40 ข้อไม่พบอัตราความคลาดเคลื่อนชนิดที่ 1 แต่ยังคงมีการประมาณค่าผิดพลาด ส่วนค่า χ^2_{MH} ระหว่างแบบสอบทั้งสองขนาดไม่แตกต่างกันอย่างมีนัยสำคัญ จึงได้เสนอว่าในการตรวจสอบการทำหน้าที่ต่างกันต้องพิจารณาทั้ง α_{MH} และ χ^2_{MH}

ภายใต้เงื่อนไขที่ทำหน้าที่ต่างกันพบว่าค่าอำนาจจำแนก ค่าความยาก ค่าการเดา การกระจายความสามารถ และปฏิสัมพันธ์ระหว่างการกระจายความสามารถกับค่าพารามิเตอร์ของแบบสอบ (a,b,c) มีผลต่อการประมาณค่า α_{MH} แต่ไม่มีผลต่ออัตราความคลาดเคลื่อนชนิดที่ 2 โดยพบว่าการประมาณค่า α_{MH} ผิดพลาดในแบบสอบทั้งสองขนาด และค่า α_{MH} ที่ได้จากแบบสอบทั้งสองขนาดแตกต่างอย่างมีนัยสำคัญ ข้อสอบที่ตรวจพบการทำหน้าที่ต่างกันส่วนใหญ่เป็นข้อสอบที่ทำหน้าที่ต่างกันแบบเอกรูปมากกว่าแบบอเนกรูป

Mazor และ Clauser (1995) ศึกษาเปรียบเทียบวิธีถดถอยโลจิสติกกับวิธีแมนเทล - แฮนส์เซลในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบกรณีที่ใช้เกณฑ์ภายนอกเป็นความสามารถหลากหลาย (multiple ability) ศึกษาจากข้อมูลจริง จากผลการตอบแบบสอบวัดผลสัมฤทธิ์ทางการเรียนวิชาประวัติศาสตร์และวิชาเคมีของนักเรียนระดับมัธยมศึกษา โดยกลุ่มผู้สอบจำแนกตามเพศ ความสามารถทางภาษา ส่วนเกณฑ์ที่ใช้ตรวจสอบแปรเปลี่ยนดังนี้ 1) ใช้คะแนนรวมของแบบสอบตรวจสอบ 2) ใช้คะแนนรวมของแบบสอบและตัวแปรความสามารถภายนอก ในที่นี้ใช้ความถนัดทางภาษาและความถนัดทางคณิตศาสตร์เป็นตัวแปรที่เพิ่มเข้ามา ส่วนความยาวแบบสอบใช้ 75 ข้อในทุกเงื่อนไข

ผลการศึกษาพบว่าวิธีถดถอยโลจิสติกตรวจพบข้อสอบที่ทำหน้าต่างกันได้มากกว่าวิธีแมนเทล-แฮนส์เซลในทุกเงื่อนไข แต่ผลที่ได้จากการตรวจสอบสอดคล้องกันกล่าวคือ เมื่อนำคะแนนความถนัดทางภาษาและคะแนนความถนัดทางคณิตศาสตร์ เข้ามาเป็นเกณฑ์รวมกับคะแนนรวมของแบบสอบจะทำให้ตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ลดลง โดยคะแนนความถนัดทางภาษาเมื่อรวมกับคะแนนรวมของแบบสอบทำให้ตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ต่ำกว่าเมื่อใช้คะแนนความถนัดทางคณิตศาสตร์รวมกับคะแนนรวมของแบบสอบ Clauser ได้อธิบายเพิ่มเติมว่าความสามารถที่เพิ่มเข้ามาหากมีความตรงและมีความสัมพันธ์กับแบบสอบและทำให้ข้อสอบข้อนั้นทำหน้าที่ต่างกันควรจะตัดข้อสอบข้อนั้นทิ้ง ซึ่งข้อสอบลักษณะนี้มีลักษณะเป็นข้อสอบพหุมิติที่เกิดจากความแตกต่างของความสามารถที่สองที่มีความสัมพันธ์กับคุณลักษณะแฝงหลัก ดังนั้นหากตรวจสอบข้อสอบพหุมิติที่ทำหน้าที่ต่างกัน ควรนำเกณฑ์ความสามารถมาคิดคำนวณด้วยเพราะทำให้ผลการตอบมีความถูกต้องมากขึ้นและวิธีที่มีความเหมาะสมในการตรวจสอบข้อสอบลักษณะเช่นนี้คือวิธีถดถอยโลจิสติก

Roussos และ Stout (1996) ศึกษาผลกระทบของกลุ่มตัวอย่างขนาดเล็กที่มีต่ออัตราความคลาดเคลื่อนชนิดที่ 1 ของวิธีซิบเทสท์กับวิธีแมนเทล-แฮนส์เซล ศึกษา 2 ครั้ง ครั้งแรกใช้ขนาดกลุ่มตัวอย่าง 100 200 300 และ 1000 คน ความแตกต่างของค่าเฉลี่ยการกระจายความสามารถระหว่างกลุ่มเป็น 0.0 0.5 และ 1.0 ใช้ความยาวแบบสอบจำนวน 25 ข้อ ผลพบว่าค่าสถิติของวิธีซิบเทสท์กับวิธีแมนเทล-แฮนส์เซลมีแนวโน้มที่จะมีอัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้นเมื่อขนาดกลุ่มตัวอย่างและความแตกต่างของค่าเฉลี่ยความสามารถระหว่างกลุ่มเพิ่มขึ้น

การศึกษาครั้งที่ 2 ใช้ขนาดกลุ่มตัวอย่าง 500 1000 และ 3000 คน ความแตกต่างของค่าเฉลี่ยการการจายความสามารถระหว่างกลุ่มเป็น 0.0 และ 1.0 ค่าอำนาจจำแนก 3 ระดับ ค่าความยาก 5 ระดับ ค่าการเดา 3 ระดับ ผลการศึกษาพบว่าเมื่อมีความแตกต่างของค่าเฉลี่ยการกระจายความสามารถเป็น 1.0 จะทำให้อัตราความคลาดเคลื่อนชนิดที่ 1 เพิ่มขึ้นมากทั้งสองวิธี โดยวิธีแมนเทล-แฮนส์เซลมีการระบุผิดพลาดมากกว่าวิธีซิบเทสท์ ส่วนข้อสอบที่มีการระบุผิดพลาดส่วนใหญ่เป็นข้อสอบที่มีค่าอำนาจจำแนกสูงและค่าความยากต่ำในทุกขนาดกลุ่มตัวอย่าง เมื่อเพิ่มขนาดกลุ่มตัวอย่างเป็น 3000 คน อัตราความคลาดเคลื่อนของทั้งสองวิธีมีแนวโน้มสูงขึ้นและเมื่อไม่มีความแตกต่างของค่าเฉลี่ยการกระจายความสามารถทั้งสองวิธีให้ผลที่น่าพอใจในทุกเงื่อนไข

Narayanan และ Swaminathan (1996) เปรียบเทียบประสิทธิภาพระหว่างวิธีแมนเทล-แฮนส์เซล วิธีถดถอยโลจิสติกและวิธีโครชิบท์ (CRO-SIB) ในการตรวจสอบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป ศึกษาจากข้อมูลจำลอง ภายใต้ปัจจัยที่แปรเปลี่ยนดังนี้ 1) กลุ่มตัวอย่าง ใช้กลุ่มตัวอย่าง 500 1000 คนในกลุ่มอ้างอิง และ 200 500 คนในกลุ่มเปรียบเทียบ 2) การกระจายความสามารถระหว่างกลุ่มอ้างอิงและกลุ่มเปรียบเทียบเท่ากันกับไม่เท่ากัน 3) สัดส่วนของข้อสอบที่ทำหน้าที่ต่างกันในแบบสอบโดยจำลองข้อมูลใน 3 ระดับคือ 0% 10% และ 20% 4) ขนาดของข้อสอบที่ทำหน้าที่ต่างกันหรือพื้นที่ความแตกต่างระหว่างโค้งคุณลักษณะข้อสอบสองกลุ่มกำหนด 4 ระดับคือ 0.4 0.6 0.8 และ 1.0 5) ค่าความยากและค่าอำนาจจำแนกของข้อสอบส่วนค่าการเดากำหนดให้เท่ากันคือ 0.2 และศึกษาความยาว 40 ข้อในทุกเงื่อนไข

ผลการศึกษาพบว่าวิธีถดถอยโลจิสติกและวิธีโครชิบท์ให้ผลใกล้เคียงกันในการตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป และทั้งสองวิธีตรวจสอบได้ดีกว่าวิธีแมนเทล-แฮนส์เซล ปัจจัยที่มีผลต่อการตรวจสอบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปมีดังนี้ ขนาดกลุ่มตัวอย่าง เมื่อเพิ่มขนาดกลุ่มตัวอย่างทั้งสามวิธีสามารถตรวจสอบได้เพิ่มขึ้น การกระจายความสามารถระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบเมื่อกำหนดให้เท่ากันจะทำให้อัตราการตรวจสอบสูงขึ้น และพื้นที่ความแตกต่างระหว่างกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบเพิ่มขึ้นจาก 0.4 เป็น 1.0 ทั้งสามวิธีสามารถตรวจสอบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปได้เพิ่มขึ้น ข้อสอบที่ตรวจพบว่าทำหน้าที่ต่างกันด้วยวิธีถดถอยโลจิสติกและวิธีโครชิบท์ส่วนใหญ่เป็นข้อสอบที่ค่าความยากต่ำ ค่าอำนาจจำแนกสูง ส่วนวิธีแมนเทล-แฮนส์เซลตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ดีเฉพาะกรณีข้อสอบยากและข้อสอบง่ายซึ่งเป็นลักษณะที่โค้งคุณลักษณะข้อสอบของผู้สอบสองกลุ่มตัดกันที่ความสามารถสูงหรือความสามารถต่ำ เท่านั้น

สำหรับการตรวจค้นผิดพลาดประเภทที่ 1 พบว่าวิธีแมนเทลแฮนส์เซลมีอัตราการตรวจค้นผิดพลาดต่ำกว่าวิธีถดถอยโลจิสติกและวิธีโครชิบท์ วิธีโครชิบท์ตรวจค้นผิดพลาดสูงสุด

กาญจนา วัธนสุนทร (2537) ได้พัฒนาเกณฑ์ตัดสินข้อสอบลำเอียงทางเพศด้วยข้อมูลเชิงประจักษ์ ดัชนีที่พัฒนาขึ้นได้แก่ พื้นที่ระหว่างดค้งการตอบข้อสอบชนิดคิดเครื่องหมาย (SA) และไม่คิดเครื่องหมาย (UA)จากวิธีทฤษฎีการตอบสนองข้อสอบ ดัชนี α_{MH} จากวิธีแมนเทล-แฮนส์เซล และดัชนี β_{SIB}จากวิธีชิบเทสท์ โดยใช้ข้อมูลการตอบข้อสอบคัดเลือกเข้าศึกษาต่อในสถาบันอุดมศึกษา ปีการศึกษา 2535 ความยาวแบบสอบ 20 30 และ 40 ข้อสำหรับวิชาคณิตศาสตร์ และ 50

60 70 และ 80 ข้อสำหรับวิชาภาษาอังกฤษ และใช้ขนาดผู้สอบ 100 200 400 600 800 และ 1000 คน

ผลการวิจัยพบว่าเกณฑ์ที่พัฒนาขึ้นเพื่อใช้ตัดสินความลำเอียงข้อสอบมีดังนี้

1) |SA| > .80 และ UA > .50 เมื่อความยาวแบบสอบต่ำกว่า 50 ข้อ

2) |SA| > .40 และ UA > 1.20เมื่อความยาวแบบสอบ 50 ข้อขึ้นไป

3) .60 > α_{MH} > 1.40 และ β_{SIB} > .06 ทุกขนาดผู้สอบและความยาวแบบสอบ

ทั้งนี้ในการใช้ดัชนี SA หรือ UA ควรใช้ผู้สอบขนาด 800 คนขึ้นไป ส่วนดัชนี α_{MH} และ β_{SIB} ควรใช้ขนาดผู้สอบอย่างน้อย 600 คน

ผลการตรวจค้นข้อสอบลำเอียงทางเพศ เมื่อใช้ดัชนีตามที่กำหนดพบว่า มีความไม่คงที่ข้ามขนาดผู้สอบและความยาวแบบสอบ ความสอดคล้องในการตรวจค้นข้อสอบลำเอียงภายในวิธีเดียวกันข้ามขนาดผู้สอบค่อนข้างต่ำ แต่จะสูงขึ้นเมื่อขนาดผู้สอบ 600 คนขึ้นไป สำหรับการวิเคราะห์ความลำเอียงของข้อสอบที่มีต่อเพศผู้สอบพบว่า ข้อสอบลำเอียงเข้าข้างผู้หญิงในวิชาภาษาอังกฤษ ส่วนวิชาคณิตศาสตร์ข้อสอบลำเอียงที่พบลำเอียงเข้าข้างเพศชาย เมื่อใช้ดัชนี SA และ α_{MH} ในขณะที่ดัชนี β_{SIB} ให้ผลตรงข้าม

เกษร หว่างจิตร (2539) ศึกษาการทำหน้าที่ต่างกันของข้อสอบวิชาภาษาไทยและภาษาอังกฤษด้วยวิธีแมนเทล-แฮนส์เซล โดยกลุ่มผู้สอบจำแนกตามเพศ ภูมิลำเนา ประสบการณ์ในการสอบและสังกัดสถานศึกษา ข้อมูลที่ใช้เป็นผลการตอบข้อสอบวิชาสอบร่วมในส่วนที่เป็นข้อสอบแบบเลือกตอบของศูนย์ทดสอบทางภาษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย รหัสการสอบ 383 กลุ่มตัวอย่างที่ใช้เป็นผู้สอบวิชาภาษาไทย 506 คน และวิชาภาษาอังกฤษ 501 คน

ผลการศึกษาพบว่าข้อสอบที่ทำหน้าต่างกันส่วนใหญ่เป็นแบบอเนกรูป เมื่อวิเคราะห์กลุ่มผู้สอบจำแนกตามตามเพศ จะพบข้อสอบที่ทำหน้าที่ต่างกันทั้งแบบเอกรูปและอเนกรูปมากที่สุด รองลงมาคือการจำแนกตามภูมิลำเนา สังกัดสถานศึกษา และประสบการณ์ในการสอบตามลำดับ เมื่อพิจารณาข้อสอบที่ทำหน้าต่างกันพบว่าส่วนมากเป็นข้อสอบที่มีค่าอำนาจจำแนกค่อนข้างต่ำทั้งสองวิชา สำหรับข้อสอบที่ทำหน้าที่ต่างกันในวิชาภาษาไทยส่วนใหญ่เป็นข้อสอบที่ง่าย ส่วนวิชาภาษาอังกฤษข้อสอบที่ทำหน้าที่ต่างกันเป็นข้อสอบที่ยากมาก

จิติมา วรรณศรี (2540) ได้เปรียบเทียบประสิทธิภาพในการตรวจสอบการทำหน้าที่ต่างกันระหว่างวิธีแมนเทล-แฮนส์เซลกับวิธีซิบเทสท์ โดยศึกษาจากข้อมูลจำลอง ปัจจัยที่ศึกษาได้แก่

ความยาวแบบสอบ 3 ขนาด คือ 30 60 และ 90 ข้อ ขนาดกลุ่มตัวอย่าง 3 ขนาด คือ 200 600 และ 1000 คน โดยแต่ละขนาดมีอัตราส่วนระหว่างผู้สอบกลุ่มอ้างอิงกับกลุ่มเปรียบเทียบต่างกัน คือ 1:1 1:0.9 1:0.75 และ 1 : 0.5

ผลการศึกษาพบว่าวิธีแมนเทล-แฮนส์เซลกับวิธีซิบเทสท์ มีประสิทธิภาพเท่าเทียมกันในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบที่ทุกขนาดกลุ่มตัวอย่างและทุกอัตราส่วนภายใต้ความยาวแบบสอบเดียวกัน และเมื่อใช้แบบสอบที่มีความยาวปานกลาง (60 ข้อ) ทั้งสองวิธีสามารถตรวจสอบได้อย่างมีประสิทธิภาพที่สุด นอกจากนี้เมื่อใช้ขนาดกลุ่มตัวอย่างมากขึ้นจะสามารถตรวจสอบข้อสอบที่ทำหน้าที่ต่างกันได้ถูกต้องมากขึ้น โดยส่วนมากวิธีซิบเทสท์ มีอัตราความคลาดเคลื่อนชนิดที่ 1 มากกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย

เสรี ชัดแช้ม (2540) ศึกษาเปรียบเทียบผลการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูประหว่างวิธีแมนเทล-แฮนส์เซลแบบปกติกับวิธีแมนเทล-แฮนส์เซลแบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ ศึกษาจากใช้ข้อมูลผลการตอบแบบสอบวัดความสามารถในการอ่านภาษาไทยของนักเรียนชั้นมัธยมศึกษาปีที่ 1 สังกัดกรมสามัญศึกษา จังหวัดชลบุรี จำนวน 1200 คน โดยกลุ่มผู้สอบจำแนกตามเพศ

ผลการศึกษาพบว่าวิธีแมนเทล-แฮนส์เซลแบบแบ่งกลุ่มความสามารถของผู้สอบและความยากของข้อสอบ สามารถตรวจพบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูปได้สอดคล้องกับวิธี IRT area และตรวจพบเพิ่มขึ้นจากวิธีแมนเทล-แฮนส์เซลแบบปกติ ข้อสอบที่ตรวจพบส่วนใหญ่เป็นข้อสอบยากง่ายปานกลาง และข้อสอบง่าย ซึ่งมีลักษณะโค้งลักษณะข้อสอบของกลุ่มผู้สอบสองกลุ่มตัดกันบริเวณใกล้ๆจุดกลางของช่วงความสามารถ

จากผลการศึกษางานวิจัยข้างต้น พบว่าในการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูปวิธีแมนเทล-แฮนส์เซลเป็นวิธีที่นิยมและใช้กันอย่างกว้างขวาง เพราะเป็นวิธีที่ไม่ยุ่งยากใช้กลุ่มตัวอย่างขนาดเล็ก คำนวณง่าย (French and Miller, 1996 ; Roussos and Stout, 1996) ผลการตรวจสอบการทำหน้าที่ต่างกันแบบเอกรูปด้วยวิธีแมนเทล-แฮนส์เซลให้ผลสอดคล้องกับวิธีทฤษฎีการตอบสนองข้อสอบ วิธีถดถอยโลจิสติก และวิธีซิบเทสท์ (Hambleton, 1989 ; Rogers and Swaminathan, 1993 ; Narayanan and Swaminathan, 1996 ; จิติมา วรรณศรี, 2540)

สำหรับข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูป โดยทั่วไปพบได้บ่อยในข้อมูลจริง (Mellenbergh, 1982 cited in Narayanan and Swaminathan, 1996 ; เกษร หว่างจิตร, 2539) ซึ่งยัง

มีการศึกษาไม่มากนัก ผลจากการศึกษางานวิจัยพบว่าวิธีที่สามารถนำมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้คือวิธีทฤษฎีการตอบสนองข้อสอบ วิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติก แต่วิธีทฤษฎีการตอบสนองข้อสอบ มีข้อยุ่งยากในการปฏิบัติ ส่วนวิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติก มีความสะดวกและง่ายต่อการปฏิบัติ (Narayanan and Swaminathan, 1996) ข้อค้นพบจากการตรวจสอบข้อสอบทำหน้าที่ต่างกันแบบอเนกรูป พบว่าวิธีแมนเทล-แฮนส์เซลมีอัตราการตรวจสอบข้อสอบทำหน้าที่ต่างกันต่ำกว่าวิธีถดถอยโลจิสติก แต่อัตราความคลาดเคลื่อนประเภทที่ 1 ของวิธีถดถอยโลจิสติกสูงกว่าวิธีแมนเทล-แฮนส์เซลเล็กน้อย (Roger and Swaminathan, 1990 ; Narayanan and Swaminathan, 1996) แต่ข้อค้นพบดังกล่าวขัดแย้งกับการศึกษาของ Mazor (1994) ที่พบว่าวิธีแมนเทล-แฮนส์เซลสามารถตรวจการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูปได้ดี ถ้าแบ่งกลุ่มผู้สอบตามคะแนนผลการสอบออกเป็นกลุ่มที่มีความสามารถต่ำและกลุ่มที่มีความสามารถสูง โดยไม่ทำให้อัตราความคลาดเคลื่อนประเภทที่ 1 สูงขึ้น ซึ่งข้อค้นพบของ Mazor (1994)สอดคล้องกับการศึกษาของ เสรี ชัดแช้ม (2540)ที่พบว่าวิธีแมนเทล-แฮนส์เซลแบบแบ่งกลุ่มความสามารถสามารถตรวจพบข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปได้ดีกว่าวิธีแมนเทล-แฮนส์เซลแบบปกติ

ปัจจัยที่มีผลต่อประสิทธิภาพการตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป คือค่าความยากและค่าอำนาจจำแนกของข้อสอบ เช่น งานวิจัยของ Rogers and Swaminathan (1993) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปที่ตรวจพบด้วยวิธีแมนเทล-แฮนส์เซลและวิธีถดถอยโลจิสติกส่วนใหญ่เป็นข้อสอบที่ยากง่ายปานกลาง ค่าอำนาจจำแนกสูง และวิธีแมนเทล-แฮนส์เซลตรวจสอบข้อสอบแบบอเนกรูปได้ดีเฉพาะกรณีข้อสอบง่ายและข้อสอบยากเท่านั้น แต่ผลการศึกษาของ Narayanan และ Swaminathan (1996) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปส่วนใหญ่เป็นข้อสอบที่มีค่าความยากต่ำ ค่าอำนาจจำแนกสูง ผลการศึกษาในประเทศไทยของเกษร หว่างจิตร (2539) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปส่วนใหญ่เป็นข้อสอบที่มีค่าอำนาจจำแนกต่ำในวิชาภาษาไทยและวิชาภาษาอังกฤษ และเป็นข้อสอบที่ง่ายมากในวิชาภาษาไทย แต่ในวิชาภาษาอังกฤษเป็นข้อสอบที่ยาก ผลการศึกษาของเสรี ชัดแช้ม (2540) พบว่าข้อสอบที่ทำหน้าที่ต่างกันแบบอเนกรูปส่วนใหญ่เป็นข้อสอบยากง่ายปานกลางและข้อสอบง่าย ดังนั้นเพื่อเป็นการยืนยันข้อค้นพบให้มีความชัดเจนมากขึ้นจึงต้องศึกษาเกี่ยวกับค่าความยากของข้อสอบและค่าอำนาจจำแนกของข้อสอบ เมื่อมีการจัดกลุ่มความสามารถของผู้สอบต่างกันว่ามีผลต่ออัตราการตรวจสอบและอัตราความความคลาดเคลื่อนของวิธีที่นำมาตรวจสอบการทำหน้าที่ต่างกันของข้อสอบแบบอเนกรูป